ARTICLE IN PRESS

International Journal of Forecasting xxx (xxxx) xxx

Contents lists available at ScienceDirect

International Journal of Forecasting

journal homepage: www.elsevier.com/locate/ijforecast

A brief history of forecasting competitions

Rob J. Hyndman

Department of Econometrics & Business Statistics, Monash University, Clayton VIC 3800, Australia

ARTICLE INFO

Keywords:
Evaluation
Forecasting accuracy
Kaggle
M competitions
Neural networks
Prediction intervals
Probability scoring
Time series

ABSTRACT

Forecasting competitions are now so widespread that it is often forgotten how controversial they were when first held, and how influential they have been over the years. I briefly review the history of forecasting competitions, and discuss what we have learned about their design and implementation, and what they can tell us about forecasting. I also provide a few suggestions for potential future competitions, and for research about forecasting based on competitions.

1. Introduction

Prediction competitions go back millennia; for example, rival diviners in ancient Greece competed to predict the future more accurately (Raphals, 2013, p. 124). However, the history for general time series forecasting (i.e., predicting the future of regularly observed data over time) is much more limited, going back only about 50 years. In fact, it wasn't until computers were widely available that it became feasible for forecasting competitions to be held at all.

Time series forecasting competitions have been a feature of the *International Journal of Forecasting* and the *Journal of Forecasting* ever since the journals were founded in the early 1980s. This strong emphasis on large-scale empirical evaluations of forecasting methods, and the need to compare newly proposed methods against existing state-of-the-art methods, has played a large part in pushing researchers to develop new methods that can be shown to work in practice (Fildes & Ord, 2002).

Researchers who are new to forecasting are often surprised to learn how controversial such competitions were when they were first conducted about 50 years ago. I review this controversy in Section 2. The influential series of Makridakis competitions are discussed in Section 3, and other forecasting competitions are described in Section 4. Finally, I provide a few comments on the future of forecasting competitions, and research about forecasting competitions, in Section 5. I do not cover forecasting competitions that are not based around time series data.

2. Early controversy

The earliest forecasting competitions were between methods rather than people. Given the communication tools available at the time, it was not feasible to conduct large-scale forecasting competitions involving many entrants spread around the world. Thus the first few competitions involved individual researchers comparing the accuracy of several methods applied to multiple time series. I only include the first two of these. From 1980 onwards, my scope is restricted to competitions involving multiple entrants.

2.1. Nottingham studies

The earliest non-trivial study of time series forecast accuracy was probably that conducted by David Reid as part of his PhD at the University of Nottingham (Reid, 1969). Building on his work, Paul Newbold and Clive Granger then conducted a study of forecast accuracy involving 106 time series (Newbold & Granger, 1974). Although they did not invite others to participate, they did start the discussion as to what forecasting methods are the most accurate for different types of time series. They presented their ideas to the Royal Statistical Society, and the subsequent

E-mail address: Rob.Hyndman@monash.edu.

https://doi.org/10.1016/j.ijforecast.2019.03.015
0169-2070/

Please cite this article as: R.J. Hyndman, A brief history of forecasting competitions. International Journal of Forecasting (2019), https://doi.org/10.1016/j.ijforecast.2019.03.015.

discussion reveals some of the erroneous thinking of the time.

One important feature of their results was the empirical demonstration that forecast combinations improve the accuracy. A similar result had been demonstrated as far back as Francis Galton in 1907 (Wallis, 2014), yet one discussant (GJA Stern) stated

> "The combined forecasting methods seem to me to be non-starters … Is a combined method not in danger of falling between two stools?"

Maurice Priestley, later to become the founding and long-serving Editor-in-Chief of the *Journal of Time Series Analysis*, said

> "The authors' suggestion about combining different forecasts is an interesting one, but its validity would seem to depend on the assumption that the model used in the Box-Jenkins approach is inadequate—for otherwise, the Box-Jenkins forecast alone would be optimal".

This reveals a view commonly held (even today) that there is some single model that describes the data generating process, and that the job of a forecaster is to find it. This seems patently absurd to me — real data come from processes that are much more complicated, non-linear and non-stationary than any model we might dream up — and George Box himself famously dismissed it saying, "All models are wrong but some are useful".

There was also a strong bias against automatic forecasting procedures. For example, Gwilym Jenkins said

> "The fact remains that model building is best done by the human brain and is inevitably an iterative process".

Perhaps Jenkins was reflecting the widely-held view that the type of intuitive thinking and extensive experience that are typically involved in model building cannot be represented by an algorithm or mathematical model. Subsequent history has shown that to be untrue provided that enough data are available and the model is flexible enough to capture the variations seen in real data.

Of course, human judgment still has value in forecasting, as was demonstrated by Petropoulos, Kourentzes, Nikolopoulos, and Siemsen (2018), who show that combining judgment with statistical models can lead to statistically significant improvements in forecast accuracy.

3. The Makridakis competitions

3.1. Makridakis and Hibon (1979)

Five years later, Spyros Makridakis and Michèle Hibon put together a collection of 111 time series and compared many more forecasting methods. They also presented the results to the Royal Statistical Society. The resulting paper (Makridakis & Hibon, 1979) seems to have caused quite a stir, and the discussion published along with the paper is entertaining, and at times somewhat shocking.

Maurice Priestley was in attendance again, and still clinging to the view that there was a true model waiting to be discovered:

> "The performance of any particular technique when applied to a particular series depends essentially on (a) the model which the series obeys; (b) our ability to identify and fit this model correctly and (c) the criterion chosen to measure the forecasting accuracy".

Makridakis and Hibon replied,

> "There is a fact that Professor Priestley must accept: empirical evidence is in *disagreement* with his theoretical arguments".

Many of the discussants seem to have been enamoured with ARIMA models.

> "It is amazing to me, however, that after all this exercise in identifying models, transforming and so on, that the autoregressive moving averages come out so badly. I wonder whether it might be partly due to the authors not using the backwards forecasting approach to obtain the initial errors". — *W.G. Gilchrist*

> "I find it hard to believe that Box-Jenkins, if properly applied, can actually be worse than so many of the simple methods". — *Chris Chatfield*

At times, the discussion degenerated to questioning the competency of the authors:

> "Why do empirical studies sometimes give different answers? It may depend on the selected sample of time series, but I suspect it is more likely to depend on the skill of the analyst … these authors are more at home with simple procedures than with Box-Jenkins". — *Chris Chatfield*

Again, Makridakis & Hibon responded:

> "Dr Chatfield expresses some personal views about the first author … It might be useful for Dr Chatfield to read some of the psychological literature quoted in the main paper, and he can then learn a little more about biases and how they affect prior probabilities".

3.2. M-competition

In response to the hostility and the charge of incompetence, Makridakis and Hibon followed up with a new competition involving 1001 series. This time, anyone could submit forecasts, making this the first true forecasting competition (where multiple people could submit entries) as far as I am aware. They also used multiple forecast measures to determine the most accurate method.

The 1001 time series were taken from demography, industry and economics, and ranged in length between 9 and 132 observations. All of the data were either non-seasonal (e.g., annual), quarterly or monthly. Curiously, all of the data were positive, which made it possible to

Please cite this article as: R.J. Hyndman, A brief history of forecasting competitions. International Journal of Forecasting (2019), https://doi.org/10.1016/j.ijforecast.2019.03.015.